Breast Cancer Prediction Model

Project Description

This project implements a complete machine learning pipeline for predicting breast cancer diagnosis (**malignant vs benign**) using quantitative characteristics of cell nuclei extracted from digitized images of fine needle aspirate (FNA) of breast masses.

The system follows best practices in artificial intelligence and healthcare-oriented machine learning, covering data preprocessing, exploration data analysis, model training, evaluation, optimization, and deployment preparation. Special emphasis is placed on medical relevance, model robustness, and recall optimization due to the critical nature of cancer diagnosis.

Faculty / University

Arab Academy for Science, Technology and Maritime Transport

Course Information

- **Course Code**: CAI3101
- **Course Name**: Introduction to Artificial Intelligence

Course Instructors

- **Dr. Mohamed Ali Abdel-Rabuh Hamouda**
- **Eng. Mohamed Moheb Abdel-Sattar Emara**

Submitted By

- **Seif Ebeid** – ID: 231014746
- **Abdullah Nagy** – ID: 231004881

---

1. Project Overview

This project presents an end-to-end machine learning system designed to assist in breast cancer diagnosis using medical imaging data. The system processes numerical features extracted from digitized cell images and applies multiple machine learning algorithms to classify tumors as benign or malignant.

Several models are trained, evaluated, and compared to identify the most accurate and reliable approach for medical decision support. The project is intended for educational and research purposes, with a strong focus on explainability, evaluation rigor, and medical safety considerations.

---

2. Dataset Information

The project uses the **Wisconsin Breast Cancer Diagnostic (WBCD)** dataset, a well-known benchmark dataset in medical machine learning.

- **Source**: Wisconsin Breast Cancer Diagnostic Dataset
- **Total Samples**: 569
- **Total Columns**: 32
    - 30 feature columns
    - 1 patient ID column
    - 1 target (diagnosis) column
- **Missing Values**: None

Target Variable

- **diagnosis** (Binary Classification)
    - **1** → Malignant (Cancer Present)
    - **0** → Benign (Non-cancerous)

Feature Description

The dataset contains **30 numerical features** derived from digitized images of breast mass cell nuclei. Each feature describes a physical or structural characteristic of the cell nucleus.

Core Cell Nucleus Characteristics (10)

1. Radius
2. Texture

3. Perimeter
4. Area
5. Smoothness
6. Compactness
7. Concavity
8. Concave Points
9. Symmetry
10. Fractal Dimension

Measurement Types (3 per characteristic)

Each characteristic is measured in three different ways:

- **Mean**: Average value across all nuclei
- **Standard Error (SE)**: Measurement variability
- **Worst**: Mean of the three largest values

Total Features:
10 characteristics × 3 measurements = **30 features**

---

3. Data Preprocessing Pipeline

3.1 Data Loading

- Load full dataset (no sampling required)
- Dataset is small, complete, and clean
- Remove unnamed or empty columns if present

3.2 Target Encoding

- Original labels: `M` (Malignant), `B` (Benign)
- Encoded labels:
  - `M → 1`
  - `B → 0`
- Numeric encoding is required for machine learning algorithms

3.3 Feature Selection

- Dropped column: `id` (patient identifier with no predictive value)
- Retained: All 30 medical measurement features
- No missing or corrupted values detected

3.4 Data Quality Summary

- **Missing Values**: 0%
- **Data Types**: All numeric (float)
- **Class Distribution**:
  - Benign: ~63%
  - Malignant: ~37%
- **Scaling**: StandardScaler applied where required

---

4. Exploratory Data Analysis (EDA)

4.1 Target Distribution

- Visualized using count plots and pie charts
- Confirms moderate class imbalance but acceptable for training
- No resampling required

4.2 Correlation Analysis

- Computed full correlation matrix
- Identified strongest correlations with diagnosis
- Most influential features:
  - `concave_points_worst`
  - `perimeter_worst`
  - `radius_worst`
- Strong correlation observed between radius, perimeter, and area features

4.3 Feature Distribution Analysis

- Compared distributions for benign vs malignant cases
- “Worst” measurement features show clear class separation
- Confirms medical relevance of extracted features

---

5. Machine Learning Models

5.1 Data Splitting

- **Train/Test Split**: 80% / 20%

- **Stratification**: Maintains class proportions
- **Random State**: 42 (reproducibility)

5.2 Feature Scaling

- **StandardScaler** applied to:
  - Logistic Regression
  - K-Nearest Neighbors
  - Voting Ensemble
- Tree-based models are not scaled (scale-invariant)

5.3 Models Implemented

1. Logistic Regression

- Linear baseline model
- Fast and interpretable
- Used as reference model

2. Decision Tree

- Captures non-linear patterns
- Regularized to prevent overfitting
- Highly interpretable

3. Random Forest

- Ensemble of decision trees
- Reduces variance and overfitting
- Strong baseline performance

4. K-Nearest Neighbors (KNN)

- Instance-based learning
- Requires scaled features
- Captures local similarity patterns

5. Gradient Boosting

- Sequential ensemble learning
- Strong non-linear modeling
- Best overall performance

6. Tuned Gradient Boosting

- Optimized using RandomizedSearchCV
- F1-score used as tuning metric
- Final production-ready model

7. Voting Ensemble

- Combines:
  - Gradient Boosting (50%)
  - Random Forest (25%)
  - Logistic Regression (25%)
- Improves robustness and stability

---

6. Evaluation Strategy

- Train/Test evaluation
- 5-Fold and 10-Fold Cross-Validation
- Confusion Matrix analysis
- ROC Curve and AUC
- Learning curve inspection

Key Medical Priority

Recall is the most critical metric, as false negatives (missed cancer cases) are medically dangerous.

---

7. Performance Summary

- **Accuracy**: 95–98%
- **Precision**: 93–97%
- **Recall**: 94–98%
- **F1-Score**: 94–97%
- **ROC-AUC**: 0.98–0.99

Results are stable across cross-validation folds with low variance.

---

8. Feature Importance Analysis

- Extracted from Gradient Boosting model
- Most important features:
  - Worst radius, perimeter, area
  - Concave points
  - Texture features
- Align with known medical indicators of malignancy

---

9. Deployment Preparation

- Model and scaler saved using `joblib`
- Input validation enforced
- Threshold adjustable to favor recall
- Designed for integration into decision-support systems

---

10. Reproducibility

- Fixed random seeds
- Documented dependencies
- Lightweight hardware requirements
- Training time under one minute
